Talk Science to Me #65: Wie entscheidet KI?

11.6.2026 - Birgit Baustädter | Technische Universität Graz

Weißt du, wie künstliche Intelligenz entscheidet? Nein? Dann geht es dir wie fast allen Menschen. Denn künstliche Intelligenz ist wenig durchschaubar. Bettina Könighofer möchte KI erklärbar machen.

Dieser Text ist ein Transkript der Podcast-Folge und wurde im Sinne der Verständlichkeit leicht angepasst.

Weißt du, wie künstliche Intelligenz entscheidet? Warum ein autonomes Fahrzeug gerade bremst? Warum ein Large Language Model genau diese Antwort gibt? Nein? Dann geht es dir wie eigentlich allen Menschen. Denn künstliche Intelligenz ist wenig durchschaubar. Bettina Könighofer arbeitet in diesem Bereich und möchte künstliche Intelligenz schnell, aber gleichzeitig sicher und erklärbar machen. Wie sie das umsetzt, darüber hat sie mit mir im Interview gesprochen. Mein Name ist Birgit Baustädter und ihr hört Talk Science to Me, den Wissenschafts-Podcast der TU Graz.

Liebe Bettina, vielen Dank, dass du heute hier bist und mit mir über deine Forschung sprechen wirst. Du arbeitest im Bereich Bilateral AI, das ist ein sehr neuer Begriff. Was ist das genau?

Bettina Könighofer: Vielen Dank für die Einladung. Also so neu ist der Begriff gar nicht. Man hat nur vor zehn Jahren noch Symbolic AI dazu gesagt. Und der Begriff Bilateral AI ist relativ neu. Man sagt auch Hybride AI dazu. Aber es geht im Prinzip immer darum, zwei Stränge von AI miteinander zu kombinieren. Der eine ist der symbolische AI-Strang, wo es um Logik geht, um genaue Berechnungen, um Verifikation. Und der andere ist der Data-Driven-AI-Zweig, wo man aus Daten versucht, Sachen zu generalisieren und zu lernen.

Kannst du mal Beispiele dafür sagen?

Könighofer: Sehr gern. Also zuerst für Data-Driven AI, die im Moment wahrscheinlich alle Statistik verwenden. Aber auch ganz viel zum Beispiel in der Krebserkennung. Wenn du ein neuronales Netz trainierst, das dann klassifiziert hat, ob der Patient Krebs hat oder nicht. Die Art von AI, die ich mir anschaue, ist Reinforcement Learning, also verstärkendes Lernen, wo es dabei geht, Kontrollprobleme zu lösen. Also für Robotik oder für autonomes Fahren, also immer wenn es darum geht, viele Entscheidungen zu treffen, die dann einen Task lösen sollen, eine Aufgabe lösen sollen, das nennt man dann verstärkendes Lernen. Die symbolische AI ist für mich eigentlich Logik, also wenn man logische Schlüsse ziehen muss. Das hat man ja ganz oft, dass man sagt, wie trifft jetzt ein autonomes System Entscheidungen, wenn keine Entscheidung ideal ist? Welche internen logischen Schlussfolgerungen macht es, dass man den Entscheidungen auch vertrauen kann? Und das ist der Bereich von Logik. Und wenn man das miteinander kombiniert, also die extreme Leistungsfähigkeit von den Data-Driven Approaches mit den logikbasierten Ansätzen, dann kriegt man hoffentlich Systeme raus, denen man sowohl vertrauen kann, die aber auch sehr leistungsfähig sind.

Was ist also die Vision dahinter? Warum macht man das genau?

Könighofer: Also generell sind diese Machine Learning Systeme irrsinnig kompliziert und komplex. Und die durchzutesten, also strukturiert durchzutesten, wirklich alle Fälle abdecken, geht gar nicht. Auch die Umgebungen, in denen diese Systeme operieren, sind extrem komplex. Also wenn wir uns Autonomes Fahren denken, wie viele Fußgänger*innen es gibt, Fahrradfahrer*innen und anderen

Verkehrsteilnehmer*innen, das kann man gar nicht alles durchtesten. Und darum ist es wichtig, dass irgendein Teil von dem System natürlich auch erklärbar ist, verständlich ist und auf ganz klaren Regeln basiert und nicht nur alles ein trainiertes neuronales Netz ist.

Woran arbeitest du aktuell gerade? Also was ist gerade so das Thema, das sich bei euch stellt?

Könighofer: Wir arbeiten im Moment an verschiedenen Bereichen. Der eine Bereich ist, wie kann man Unsicherheiten, die ein neuronales Netzwerk in der Objektdetektion hat, in die Planung integrieren kann. Also wenn man jetzt zertifizierte Planung hat, dass das Auto sicher fährt, wie kann diese berücksichtigen, dass das neuronale Netzwerk bei der Object Detection einen Fehler gemacht hat. Was wir uns auch ansehen, ist, wie man autonomen Agents jetzt gesetzliche Regelungen beibringt. Diese Regelungen sind extrem schwer, sogar für uns Menschen zu verstehen, sonst hätten wir keine Anwälte und so weiter. Und diese ganzen Regeln, die jetzt verschiedene Prioritäten haben, sich widersprechen können, wo es Ausnahmen gibt und so weiter einem autonomen System beizubringen ist irrsinnig schwer. Was wir uns noch gerade anschauen, was ein zentraler Punkt ist, ist, wie kann man von schon gelernten Systemen Modelle ableiten, die man verstehen kann. Also wenn du jetzt ein superkomplexes System trainiert hast, das schon toll funktioniert, wie kannst du daraus dann ein kleines Modell ableiten, mit dem man interagieren kann und anhand dessen man die Entscheidung dann nachvollziehen kann?

Hat man kleine Modelle?

Könighofer: Also diese Modelle, die wir ableiten, sind dann symbolische Modelle, wo wirklich jeder Zustand eine Bedeutung hat. Jeder Zustand spiegelt Informationen der Umgebung wider, zum Beispiel das Auto ist gerade so schnell unterwegs, ist auf dieser Position, die Fußgänger*innen verhalten sich so und so. Und anhand dieser Modelle, wo dann die Zustände wirklich Informationen haben, die man als Mensch verstehen kann, interpretieren kann, kann man dann auch sehen, wie entwickelt sich das Ganze weiter und warum macht die Entscheidung Sinn, die das autonome System gerade trifft.

Wenn du sagst Sinn machen, für wen Sinn macht? Also geht es darum auch, dass man das als User*in dann versteht?

Könighofer: Also es hängt natürlich von der Anwendung ab. Wenn du jetzt zum Beispiel in deinem Auto sitzt, hast du eine gewisse Vorstellung, wie das Auto fahren sollte, weil du weißt auch, wie die Verkehrsregeln funktionieren, du weißt, was sicher und was unsicher ist. Und natürlich soll das System diese Regeln einhalten. Im Allgemeinen, also bei uns funktioniert das so, wir haben formale Spezifikationen vorgegeben und diese formale Spezifikationen sind in irgendeiner Art Logik formuliert. Und wir überprüfen dann, dass diese formalen Spezifikationen immer eingehalten werden. Also das System, das AI-System darf nur Entscheidungen treffen, die diese Regeln akzeptiert.

Welche Ansätze habt ihr, um diese beiden Stränge von KI zusammenzubringen? Gibt es da jetzt schon Möglichkeiten oder sind das alles noch Zukunftsthemen?

Könighofer: Natürlich kombinieren wir diese Methoden schon. Nicht nur wir, sondern ganz viele Forscher*innen in Österreich. Und es gibt verschiedene Arten und Weisen, wie man das machen kann. Die Methode, die wir hauptsächlich in Graz verwenden, ist, das wirklich während der Laufzeit durchzuführen. Und zwar, dass wir im Parallelen zum AI-System ein logisches System haben. Wir nennen das typischerweise ein Schild. Und dieses Schild führt eine Sicherheitsberechnung durch für jede Aktion, die das AI-System durchführen möchte. Und nur wenn diese Sicherheitsaktion durchgeht, dann darf das AI-System diese Aktion ausführen. Es gibt aber natürlich auch andere

Ansätze. Ein anderer Ansatz, wie wir das Ganze machen, ist, während dem Trainieren, während dem Lernen des AI-Systems die Reward Function zu verändern. Das Ganze funktioniert so, dass das AI-System lernt, indem es immer Aktionen versucht, ausprobiert und dann die Umgebung beobachtet. In welchem Zustand ist jetzt die Umgebung, wo war sie vorher, wo ist sie jetzt und wie gut hat die Aktion funktioniert und hat mich das näher gebracht an das, was ich eigentlich lösen wollte oder war es vielleicht sogar schädlich, was ich gemacht habe. Und das ist dieses Reward-Signal. Also es kriegt immer Feedback, entweder positiv oder negativ. Und jetzt, wenn man so ein externes, logisches Modul hat,was die Sicherheit der Aktionen berechnen kann, überprüfen kann, kann das natürlich diese Reward Function auch mit beeinflussen und sagen, der Aktion gebe ich jetzt gleich einmal einen fetten Minus 10er, weil die ist mir zu unsicher. Und so kann man das auch beim Lernen schon mitgeben, diese Kombination von Risikoberechnung, die das logische Modul macht, und dem Ausprobieren, das das Machine Learning Modul macht.

Du hast jetzt schon erwähnt, es gibt diese Möglichkeiten, wie man diese beiden Stränge zusammenbringt. Was muss jetzt noch gemacht werden? Also was ist die Vision in der Forschung, wo man hin möchte?

Könighofer: Also ich glaube, eine große Vision von diesem Bilaterale AI-Projekt, das wir jetzt in Österreich haben, ist wirklich diese Techniken in die Praxis zu bringen. Ein ganz großer Fokus ist, Industriepartner*innen einzubinden.Es gibt auch eigene Ausschreibungen von FFG, so dass wir jetzt wirklich explorieren, wie bringen wir diese Methoden in die Praxis.

Welche Hürden gibt es da bis jetzt?

Könighofer: Also es gibt mehrere Herausforderungen, die es zu Überwinden gilt. Eine davon ist Skalierbarkeit. Also diese ganzen logischen Methoden müssen schnell sein, wenn man sich überlegt, wie toll die Anwendungen von Machine Learning sind, was die schon alles lösen können, da müssen diese logischen Methoden mithalten können und schnell genug die Berechnungen machen. Das andere, was auch ist, ist, diese logischen Methoden musst du vorher modellieren, damit du das dann logisch berechnen kannst. Und das ist natürlich auch ein Zeitaufwand. Das heißt, ein Engineering-Aufwand, man muss die Firmenpartner*innen eng mit einbeziehen und das ist natürlich auch Zeitaufwand für die Firmenpartner*innen, um dann die Modellierungen und die Systeme richtig umzusetzen zu können und dann halt gemeinsam zu integrieren.

Wie bist du eigentlich in diesen Forschungsbereich gekommen? Was hat dich da so interessiert daran? Oder tut es heute noch? Was interessiert dich heute noch so dran?

Könighofer: Ich habe mein Doktorat angefangen in formale Methoden und in formale Methoden ist es darum gegangen, wie können wir Software-Systeme überprüfen, verifizieren, dass die richtig sind. Und dann, ungefähr 2017, 2016 herum, kam dieser große Machine-Learning-Boom und dann hat sich auf einmal die Frage gestellt, okay, wie verifizieren wir jetzt diese Machine-Learning-Systeme? Und das war dann eigentlich damals total interessant und wir waren eine der ersten Veröffentlichungen damals, die eine Methode vorgestellt hat, wie man beweisen kann, dass ein Machinen-Learning-System richtig funktioniert. Das hat mir Spaß gemacht und dann sind wir auf dem Bereich geblieben und forschen noch heute dran.

Das war damals schon dieses Schild, oder?

Könighofer: Ja, genau.

Und wenn man jetzt weiter zurückgeht, also wie bist du in die Informatik gekommen und wie bist du in die formalen Methoden gekommen?

Könighofer: Also ich habe schon die HTL für Elektrotechnik besucht. Das heißt Informatik war dann nicht mehr so weit weg. Da war dann eher nur ein bisschen ein enger Spielraum zwischen mache ich jetzt Telematik oder Informatik oder Elektrotechnik oder Mathematik. Aber, dass es irgendwas in dem Viererbereich war, war dann schon sehr klar. Dass es dann so theoretisch geworden ist, habe ich mir am Anfang nicht gedacht. Das ist dann mit einer spannenden Masterarbeit passiert, die in ein Doktorat übergeführt hat. Und ich bin sehr froh, dass ich noch in dem Bereich forschen darf.

Was ist so dein Zukunftswunsch? Also gibt es irgendwo eine Forschungsfrage, die du total gern beantworten möchtest im Laufe deiner Karriere?

Könighofer: Ich glaube, so der Endgegner ist wirklich, dass wir sagen, wir haben autonome Systeme, denen man wirklich vertrauen kann, wo wir uns darauf verlassen können, dass die ethisch richtige Entscheidungen treffen.

Danke für das Interview.

Könighofer: Vielen Dank.

Vielen Dank, dass ihr in dieser Staffel wieder mit dabei wart. Wir hören uns.

https://www.tugraz.at/news/artikel/talk-science-to-me-65-wie-entscheidet-ki